The Report Committee for Radhika Sundar
Certifies that this is the approved version of the following report:

Bayesian Inference for Random Partitions

APPROVED BY

SUPERVISING COMMITTEE:

Peter Müller, Supervisor

S. Natasha Beretvas

Bayesian Inference for Random Partitions

by

Radhika Sundar, B.A;M.A

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN STATISTICS

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2013

Bayesian Inference for Random Partitions

Radhika Sundar, M.S.Stat.
The University of Texas at Austin, 2013

Supervisor: Peter Muller

I consider statistical inference for clustering, that is the arrangement of experimental units in homogeneous groups. In particular, I discuss clustering for multivariate binary outcomes. Binary data is not very informative, making it less meaningful to proceed with traditional (deterministic) clustering methods. Meaningful inference needs to account for and report the considerable uncertainty related with any reported cluster arrangement. I review and implement an approach that was proposed in the recent literature.

Table of Contents

List of Figures

1 Introduction

1.1 Clustering on the Basis of Binary Attribute Vectors

For many real-world problems, the data we have available is often in a 'yes-no' or '0/1', that is, binary format.There are many settings in which such data arise, one of them being health care, where symptoms of disease, or other factors associated with a medical condition are recorded as present or absent.One problem of interest is finding out patterns of occurrence of symptoms of disease in different subpopulations. This can reveal insights about the nature of the disease itself. Similarly, it is also of interest to know about diseases that occur in conjunction with each other. The underlying data in such problems is in binary format.

The problem setting I consider is similar to that considered by Peter Hoff in his paper on binary clustering of cancer patients' genomic data. Consider that we have a group of patients, each of whom possesses some combination of symptoms such as flu, cough, ulcers etc. How can we group them according to the type of disease indicated by their symptoms? We would like to say with some degree of certainty that we expect that a patient with cough and flu only but none of the other symptoms, to have disease A, distinct from any other disease resulting from a different combination of the initial list of symptoms. The statistical problem this situation gives rise to is that of clustering the patients into groups, where each group is characterized by the presence or absence of certain given symptoms. Since our data just consist of 0's and 1's, this problem is not easily dealt with using classical or deterministic methods of clustering. Under this approach, clusters are formed based

on the distance between patients measured on a predetermined set of attributes. The distance metric used is Euclidean distance when we have continuous numeric data. When we have binary data, we have to use other metrics, such as the Taxi cab metric. The Bayesian approach to the problem provides more flexibility. We do not need to use any distance metric, and also, each cluster has its own unique set of attributes that can be distinct from other clusters'. One such method is outlined in Peter Hoff's paper on Bayesian clustering of genomic abnormality data on cancer patients.

1.2 Data

The original data used by Hoff was collected for 150 patients for 52 genomic locations. The data is in binary form since it records the presence or absence of abnormalities- if an abnormality is present at location j for patient n, then "1" is recorded for patient n under abnormality j. Hoff 's method clusters the genomic abnormality data into groups. Once the clusters are obtained, it is then possible to draw conclusions about patterns of occurrence of abnormalities that are likely to result in cancer.

In my implementation of Hoff's method, I worked with similar data, simulating binary data for 20 patients and 3 disease symptoms, i.e,attributes. The contribution of this report is to implement Hoff's model-based clustering for the simulated data.

2 Cluster Analysis

2.1 Deterministic clustering

Brief Overview There are a variety of cluster models which try to group data objects such that those within a given cluster are more similar to each other than objects outside the cluster , and the choice of which one to use depends on the problem.The most common models are those based on connectivity, centroid measures, statistical distributions, subspace models, density models and Graph based models. Here we discuss two commonly used techniques of clustering: Hierarchical clustering and k-means clustering.

Hierarchical clustering is a connectivity based method that groups objects that are "close" to each other to form clusters. In the R package for hierarchichal clustering, there are various types of distance functions available as well as different linkage criterions. Available distance functions are euclidean, manhattan, minkowski, canberra and binary. The linkage criterion tells how to compute distances between a data point and a cluster. Hierarchical clustering is an agglomerative method, that merges clusters successively until there is only one cluster. A dendrogram is produced which tells the user the stage at which data points(one or more) are fused into a given cluster, that is, the dendrogram gives us the hierarchy of the clusters: the user then chooses which stage of the clustering to accept as the solution. Thus, there is no one single unique partitioning of the data, but many hierarchical partitions, and the user can pick the level of the hierarchy that is the best solution.

K-means clustering is a method of clustering which computes cluster centers and assigns data points to the nearest center. The centers may or may not be data points, and an initial set of cluster centers has to be specified. The clustering is usually run several times with different initial values, since the algorithm only produces local optima. Many improvements of the basic algorithm are available which use data points as centers, pick initial centers less randomly and choose the best solution based on multiple runs.

The above discussed deterministic methods and their modifications are ideally used with numeric or categorical data. They are not recommended for binary data, since most suitable distance measures used don't provide a lot of information on how similar two data points are. Thus, a Bayesian approach seems more appropriate for clustering binary data, since we don't need any distance measures, and we can use our prior knowledge about what characteristics we expect the clusters to have. Additionally, deterministic clustering methods use a pre-specified set of attributes to group the data into clusters, whereas in Hoff's subset clustering method, different clusters can be characterized by different sets of attributes. Also, deterministic clustering requires the number of clusters to be specified as input, while Hoff's method provides us the number of clusters as part of the solution.

3 Clustering as a random quantity

Given data in the form of m-dimensional binary vectors for n patients, which tell us if the ith patient has a genomic abnormality at location j, we would like to cluster these n patient indices into k many clusters, where k is initially unknown. Unlike in the deterministic method of hierarchical clustering, in the Bayesian approach we consider the clustering arrangement of the patients as a variable and treat it as a random quantity. Thus, we may assign a prior distribution to the random clusters. Similarly, we also treat the attributes that distinguish the clusters from each other as variable, as well as introducing a random parameter that decides which attributes are to be considered as the distinguishing attributes for a given cluster. The Bayesian approach allows for a much more comprehensive model for clustering than the deterministic one, which is solely based on the distance metric measured on a fixed set of attributes.

4 A Random partition Model

To elaborate on the concept of clusters being random quantities, consider that a patient i can belong to any of $1, \ldots, K$ clusters, if we know that there are a total of K clusters. Thus, we can have a function, s defined from the set of patients $1, \ldots, n$ to the the set of clusters $1 \ldots, K$, which assigns patient i to cluster k, denoted by $s(i) = k$, or $s_i = k$. We want our model to generate an assignment of patients i to clusters k based on our prior belief about their pre-existing clustering assignment(perhaps from medical records) and the data at hand. The clusters are distinguished from each other by their values on certain specific attributes.

Our task is to identify patients into clusters using binary data on their attributes. Each cluster is characterized by the rates of certain special attributes. The non-special or background attributes in any given cluster have a standard rate that is identical for all the clusters. Each patient is assigned to a specific cluster k based on the probability contribution resulting from him having or not having attributes that are special in that cluster(the ones that are not background).

The model thus uses cluster specific parameters r_k, a relevance vector to determine which attributes are special to cluster k, and cluster specific attribute rates ${\theta^\star}_k$, which describes the distribution of the attribute rates in that cluster (We will see that ${\theta^\star}_k$ equals either ${\theta'}_k$ or θ^0 on any given attribute j). Each cluster is characterized by a unique vector ${\theta^\star}_k :,\ k = 1, \ldots, K_n)$ that represents the attribute rates for all patients in that cluster. The vector ${\theta^\star}_k$ is *cluster*-specific, so that if two patients have the same attribute rates, they will belong in the same cluster. Conversely, if two patients belong to the same cluster, we immediately know that they have the same

attribute rates.

The model also needs patient-specific parameters s which is the cluster membership function, assigning patient i to cluster k. The patient specific rate θ_{ij} is derived from the cluster k that patient i belongs to, and either equals the background rate θ^0 or the special rate ${\theta^\star}_k$ depending on which one is relevant from the corresponding cluster specific value in r_k.

The patients' data y_{ij} follow a binomial distribution with success probability of belonging to cluster k equal to ${\theta^\star}_k$. Here ${\theta^\star}_k$ equals either ${\theta'}_k$ or θ^0 based on the relevance vector r_k. We use conjugate Beta priors for both ${\theta'}_k$ and θ^0. The parameter vector r follow a Bernoulli distribution on each attribute, with success probability Pi, defined in terms of a constant λ. The cluster membership function s is given a Dirichlet prior. We start with a prior assignment of patients to clusters using s during the prior simulation. Then in the posterior simulation, we use the updated probability functions for the parameters s, r, θ', θ^0 etc to assign patients and generate attribute rates that distinguish the clusters.

Our inference goals in this procedure are: (i) to obtain the posterior distribution of the attribute rates of the clusters, and (ii) to cluster patients into groups.

4.1 Sampling Model

Our data are binary variables $y_{ij} \in \{0, 1\}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, which are indicators that tumor i has an abnormality at genomic location j. The patients are indexed by i and the attributes by j. In other words, we have n vectors, where each vector is an m-dimensional binary vector representing a given patient, with bi-

nary information on the m different attributes(abnormalities in genome locations). If two vectors y_i (i.e, two samples or patients)are in the same cluster k, then they have the same probabilities for exhibiting each of the $j = 1, \ldots, m$ attributes(here, abnormalities). The distribution of y_{ij} for a given patient i in cluster k, over all attributes $j = 1, \ldots, m$ is just the product of m Bernoulli trials with success probability equal to the respective attribute rate :

$$p(y_i \mid \theta^\star{}_k) = \prod_{j=1}^{m} (\theta^\star{}_{kj})^{y_{ij}} (1 - \theta^\star{}_{kj})^{1-y_{ij}}, k = s_i \tag{1}$$

The likelihood function for all the patients $i = 1, \ldots, n$ is then:

$$p(\mathbf{y} \mid \theta^\star{}_{\mathbf{k}}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \prod_{\mathbf{j=1}}^{\mathbf{m}} (\theta^\star{}_{\mathbf{kj}})^{\mathbf{y_{ij}}} (\mathbf{1} - \theta^\star{}_{\mathbf{kj}})^{\mathbf{1-y_{ij}}}, \mathbf{k} = \mathbf{s_i} \tag{2}$$

Letting K be the total number of clusters, the likelihood above can be rewritten as follows:

$$p(\mathbf{y} \mid \theta^\star{}_k) = \prod_{k=1}^{K} \prod_{i \in S_k} \prod_{j=1}^{m} (\theta^\star{}_{kj})^{y_{ij}} (1 - \theta^\star{}_{kj})^{1-y_{ij}} \tag{3}$$

That is, letting Y_{kj} denote the number of patients in cluster k who possess attribute j, and n_k denote the total number of patients in cluster k i.e, the size of cluster k, we get:

$$p(y_i \mid \theta^\star{}_k) = \prod_{k=1}^{K} \prod_{j=1}^{m} (\theta^\star{}_{kj})^{Y_{kj}} (1 - \theta^\star{}_{kj})^{n_k - Y_{kj}} \tag{4}$$

Above, we note that the likelihood is expressed in terms of $\theta^\star{}_{kj}$, the cluster specific attribute rates. We can define the *sample specific* attribute rate vector θ_i,

defined to equal the vector ${\theta^\star}_k$ if patient(or sample) i is in cluster k. For any attribute j for such a patient,

$$\theta_{ij} = {\theta^\star}_{kj}$$

We can have the clustering be based on different subsets of the attributes by introducing the parameter variable r_k, which is an m-dimensional binary vector, specifying which of the j attributes are relevant for a given cluster k. If an attribute j is relevant, its probability ${\theta'}_{kj}$ will be modeled by a "special" prior distribution (this is just a non-Uniform Beta(a,b)); otherwise, the probability θ^o_j of the attribute is modeled by a "baseline" distribution common to all clusters(this will just be a Uniform distribution)

Thus, we have that the probability of the attribute rate for a given cluster k is going to be one of ${\theta'}_{kj}$ or θ^o_j , according to the value of r_{kj} in the equation:

$${\theta^\star}_{kj} = r_{kj}{\theta'}_{kj} + (1 - r_{kj})\theta^o_j$$

Therefore, we see that if a given attribute j is relevant for cluster k, i.e, $r_{kj} = 1$, then

$${\theta^\star}_{kj} = {\theta'}_{kj}$$

Otherwise, if j is not relevant for cluster k, i.e, $r_{kj} = 0$, then the cluster's rate for attribute j is the background rate θ^o_j:

$${\theta^\star}_{kj} = \theta^o_j.$$

We will discuss the variable r_k further later.

4.2 Random clusters : Prior model for s

In our model, two patients are in the same cluster if their attribute rates are an exact match, for each of the attributes $j = 1, \ldots, m$. That is, the binary vectors representing the patients y_i will fall in the same cluster if the probability rates corresponding to them are an exact match. We can say that the probability that two vectors y_{i_1} and y_{i_2} coincide is the same as the probability that their attribute rates are equal, i.e., $\theta_{i_1,j} = \theta_{i_2,j}$ and they are further equal to the cluster specific rate ${\theta^\star}_{kj}$. This shows that the event $s_i = k$ is equivalent to $\theta_{ij} = {\theta^\star}_{kj}$. Let s be the cluster membership function, such that $s_i = k$ means that patient i is assigned to cluster number k. The prior distribution of s is given by

$$p(s_i = k \mid s_1, \ldots, s_{i-1}, \alpha) = \begin{cases} \frac{n_k}{i-1+\alpha} & \text{if } k = 1, \ldots, K_{i-1} \\ \frac{\alpha}{i-1+\alpha}, & \text{if } k = K_{i-1} + 1 \end{cases} \tag{5}$$

where K_{i-1} is the number of clusters among the first $(i-1)$ patients. In words, the probability that a patient is in an existing cluster k is proportional to the number of members in that cluster, while the probability of belonging in a new cluster is proportional to the parameter α.

Recall that all patients in the same cluster, say cluster k, share the same attribute rates, ${\theta^\star}_{kj}$. In other words, ${\theta^\star}_k = ({\theta^\star}_{kj},\ j = 1, \ldots, m)$ are the (common) attribute rates for all patients in cluster k. We will assume a beta distribution as a prior model for the cluster-specific attribute rates

$${\theta'}_{kj} \sim \text{Beta}(a, b)$$

and a Uniform distribution as the prior model for the background attribute rates:

$$\theta^o_j \sim U(0, 1)$$

Since ${\theta^\star}_{kj}$ can equal either ${\theta'}_{kj}$ or θ^o_j , and we have that the special rate ${\theta'}_{kj}$ is some non-Uniform Beta(a, b), and the baseline rate θ^o_{kj} is the Uniform(0,1) distribution, therefore ${\theta^\star}_{kj}$ is a draw from one of the above two distributions.

Since $s_i = k$ is equivalent to $\theta_{ij} = {\theta^\star}_{kj}$, we can write (5) alternatively as a probability model for θ_{ij}. Let $\delta_x(\cdot)$ denote a point mass at x. The prior distribution of the cluster vector s can be written as below, with π denoting the probability that the relevance parameter $r_{kj} = 1$:

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1} = \begin{cases} \delta_{{\theta^\star}_k} & \text{with prob } \frac{n_k}{i-1+\alpha},\ k = 1, \ldots, K_{i-1} \\ \pi * \text{Beta}(a, b) + (1 - \pi) * \delta_{\theta^o} & \text{with prob} \frac{\alpha}{i-1+\alpha} \end{cases}$$

4.3 Cluster-specific attribute selection

The Bayesian model allows each cluster to have its own set of distinguishing attributes, that is, its own set of attributes that have special rates. In traditional hierarchical clustering, all clusters are formed based on their values for a fixed set of previously selected attributes. However, in many applications, such as in the present case where we try to detect patterns based on a large number of different attributes, the hierarchical method does not give good results.

In Hoff's method, we introduce a parameter r_k, varying across clusters, which is an m-dimensional vector with values 0 or 1 according as whether the j-th attribute is a distinguishing or special one for the given cluster k. Here, r_{kj} itself is a random quantity, and is assigned a Bernoulli prior with a success probability(the probability that the jth attribute is special) based on a constant *lambda*. The vector r_k for cluster k is a product of m Bernoullis. If $r_{kj} = 1$ for some j and k, we know that the

jth attribute in cluster k follows the special rate ${\theta'}_{kj}$ and is relevant or *distinguishes* cluster k from other clusters. The patient specific attribute rate θ_{ij} given by

$$\theta_{ij} = r_{kj}{\theta'}_{kj} + (1 - r_{kj})\theta^o_j$$

reduces to

$$\theta_{ij} = 1 * {\theta'}_{kj}$$

4.4 Prior model for r_{kj}

The parameter that allows the different attributes $1, \ldots, j$ to be relevant(contain an abnormality or be close to one), as explained above, is r_k. The r_{kj}s can be thought of as being generated by coin flips, since they come from a product of Bernoulli distributions. We have that the prior distribution of the r_{kj} is given by

$$P(r_k = (r_{kj},\ j = 1, \ldots, m) \mid s) = \prod_{j=1}^{m} \frac{\exp(\lambda_j r_{kj})}{1 + \exp(\lambda_j)}$$

for some fixed λ_j.[That is, $p(r_{kj} = 1) = \pi$]. The product of the above expression over all clusters $k = 1, \ldots, K_n$ defines $p(r \mid s) = \prod_{k=1}^{K} p(r_k \mid s)$. Thus, for given r_{kj}, we can assign either θ^o_{kj} or ${\theta'}_{kj}$ to θ_{ij}, deterministically, depending on whether $r_{kj} = 0$ or 1. In short

$$\theta_{ij} = r_{kj}{\theta'}_{kj} + (1 - r_{kj})\theta^o_{kj}$$

where $k = s_i$.

4.5 Prior model for ${\theta'}_k$

Since the data y_{ij} follow a Bernoulli distribution, we choose a conjugate Beta(a_0, b_0) prior for the cluster specific special rates ${\theta'}_k$. For a given cluster k,

the distribution of ${\theta'}_k$ will be a product of m Beta(a_0, b_0) priors. The ${\theta'}_k$ as has already been explained, are what serve to distinguish the clusters from each other, and it is the posterior values of the ${\theta'}_k$ that we are interested in estimating.

$$p(\theta' \mid s) = p(\theta' \mid K_n) = \prod_{k=1}^{K} \prod_{j=1}^{m} {\theta'}_{kj}^{a_0 - 1} (1 - {\theta'}_{kj})^{b_0 - 1}$$

4.6 Prior model for ${\theta^0}_j$

Again since the data y_{ij} follow a Bernoulli distribution, we choose a conjugate Beta(a_0, b_0) (in our case, just the Uniform density) prior for the background rates ${\theta^0}_j$. These are identical for a given attribute j across all clusters. This is useful since different diseases may have overlapping symptoms- that is, different clusters could share attributes, and these overlapping attributes should not play a role in distinguishing the clusters. Further, as discussed earlier, unlike in the deterministic clustering case, having a model which ensures that the attributes which turn out unimportant are assigned background rates allows for distinct clusters to have distinct sets of special attributes.

$$p(\theta^o) = \prod_{j=1}^{m} \theta_j^{o(a_0 - 1)} (1 - {\theta^0}_j)^{b_0 - 1}$$

5 Joint Probability Model

The joint probability model is given by

$$P(y, r, \theta', \theta^o, s) = p(s).p(r \mid s).p(\theta' \mid s).p(\theta^o).p(y \mid s, \theta^\star)$$

which is written as

$$P(y, r, \theta', \theta^o, s) = p(s).p(r \mid s).p(\theta' \mid K_n).p(\theta^o).p(y \mid s, f(r, \theta', \theta^o))$$

We can now simulate a prior for the clusters s, for the variable r, and for the attribute rates θ_{ij} in terms of K,s, r, θ', θ^o. That is, for a given patient i, we have all the components in the equation:

$$P(s, r, \theta^o_k, \theta'_k, y) = p(s).p(r \mid s).p(\theta_i \mid r, s).p(y \mid \theta_i)$$

Explicitly writing them out, we get:

(1) $p(s)$ is the Polya urn distribution given by

$$p(s_i = k \mid s_1, \ldots, s_{i-1}, \alpha) = \begin{cases} \frac{n_k}{i-1+\alpha} & \text{if } k = 1, \ldots, K_{i-1} \\ \frac{\alpha}{i-1+\alpha}, & \text{if } k = K_{i-1} + 1 \end{cases}$$

Note $p(s_1) = 1$.

(2) $p(r \mid s)$ is the product of m Bernoullis for each of the j attributes:

$$P(r_k = (r_{kj},\ j = 1, \ldots, m) \mid s) = \prod_{j=1}^{m} Ber(r_{kj} | \frac{\exp(\lambda_j)}{1 + \exp(\lambda_j)})$$

We will henceforth denote $\frac{\exp(\lambda_j)}{1+\exp(\lambda_j)}$ as π, the success probability for the relevance vectors. In our model, we set all the λ_j to be equal to some constant value λ.

(3) $p(\theta' \mid s) = p(\theta' \mid K_n) = \text{Beta}(a_0, b_0)$

(4) $p(y \mid s, \theta^\star) = p(y \mid s, f(r, \theta', \theta^o)) = \prod_{j=1}^{m} Ber(y_{ij}|\theta_{ij}))$

Thus, the joint distribution is written as :

$$
\begin{aligned}
p(w, y) = \prod_{i=2}^{n} p(s_i|s_{1,\dots,i-1)} \prod_{k=1}^{K}\prod_{j=1}^{m} r_{kj}^{\pi}(1 - r_{kj})^{1-\pi} * \\
\prod_{k=1}^{K}\prod_{j=1}^{m} {\theta'}_{kj}^{a_0-1}(1 - {\theta'}_{kj})^{b_0-1} \prod_{j=1}^{m} {\theta^0}_{j}^{(a_0-1)}(1 - {\theta^0}_{j})^{b_0-1} * \\
\prod_{k=1}^{K}[\prod_{j:r_{kj}=1} \prod_{i \in S_k} {\theta'}_{s_i,j}^{y_{ij}}(1 - {\theta'}_{s_i,j})^{1-y_{ij}} \prod_{j:r_{kj}=0} \prod_{i \in S_k} ({\theta^0}_{j})^{y_{ij}}(1 - {\theta^0}_{j})^{(1-y_{ij})} \quad (6)
\end{aligned}
$$

6 Posterior Inference

Posterior inference in our model is not trivial since it is not easy to evaluate or sample from the joint distribution. We will have to resort to stochastic techniques.

6.1 Monte Carlo Integration and Gibbs Sampler

The joint distribution that we are interested in is not easy to evaluate,in fact, it may be impossible to evaluate analytically. Thus, we need numerical methods such as Monte Carlo integration. These are stochastic methods which evaluate an integral using its expected value over a random sample from a population which is the space of solutions for the integral.

In our case, the joint distribution is the stationary distribution of a Markov chain, constructed using the Gibbs sampler method. The Gibbs sampler is a specific type of Markov Chain. It enables us to draw from the joint distribution, by sequentially drawing from the complete conditional posterior densities of each of the parameters. In our case, we first draw from the posterior density of s, then the posterior density for the relevance indicators r_k, followed by the posterior density for the special rate ${\theta'}_k$ and then θ^0. Once a parameter is sampled, the next parameter is sampled conditioned on the latest sampled values of all the other parameters. At the end of one iteration, the vector of samples from all the conditional posterior densities of the parameters is a sample from the joint posterior distribution of all the parameters. Thus, iteratively sampling from the conditional posteriors gives us samples from the joint distribution which is hard to sample from directly.

7 Posterior functions

We start with a prior assignment of patients to clusters using s during the prior simulation. Then in the posterior simulation, we use the updated probability functions for the parameters s, r, θ', θ^0 etc to assign patients and generate attribute rates that distinguish the clusters.

The idea in posterior sampling is that the posterior density is proportional to the prior times the likelihood, in other words, the joint distribution. To determine the posterior distribution of a given parameter, we only need to know how terms depending on that parameter vary in the joint. This is explained more below.

8 Conditional Posterior model for cluster membership:s

The posterior density function for any of the parameters is proportional to the joint distribution. The terms in the joint distribution that do not depend on the parameter of interest simply become part of the constant of proportionality. Recall the joint distribution is:

$$p(w, y) = p(s)p(r|K)p(\theta'|K)p(\theta^0)p(y|\theta^\star) \tag{7}$$

that is,

$$\begin{aligned} p(w, y) = p(s_1) \prod_{i=2}^{n} p(s_i|s_{1,\ldots,i-1)} \prod_{k=1}^{K}\prod_{j=1}^{m} Ber(r_{kj}|\pi)\cdot \\ \prod_{k=1}^{K}\prod_{j=1}^{m} \text{Beta}(\theta'_{kj}|a_0, b_0) \prod_{j=1}^{m} \text{Beta}({\theta^0}_j|a_0, b_0)\cdot \\ \prod_{k=1}^{K}[\prod_{j:r_{kj}=1}\prod_{i\in S_k} Ber(y_{ij}|{\theta'}_{s_i,j})][\prod_{j:r_{kj}=0}\prod_{i\in S_k} Ber(y_{ij}|{\theta^0}_j)] \end{aligned} \tag{8}$$

which is expanded as:

$$\begin{aligned} p(w, y) = \prod_{i=2}^{n} p(s_i|s_{1,\ldots,i-1)} \prod_{k=1}^{K}\prod_{j=1}^{m} \pi_{kj}^{r} 1 - \pi^{(}1 - r_{kj})* \\ \prod_{k=1}^{K}\prod_{j=1}^{m} {\theta'}_{kj}^{a_0-1}(1 - \theta'_{kj})^{b_0-1} \prod_{j=1}^{m} {\theta^0}_j^{(a_0-1)}(1 - {\theta^0}_j)^{b_0-1}* \\ \prod_{k=1}^{K}[\prod_{j:r_{kj}=1}\prod_{i\in S_k} {\theta'}_{s_i,j}^{y_{ij}}(1 - {\theta'}_{s_i,j})^{1-y_{ij}} \prod_{j:r_{kj}=0}\prod_{i\in S_k} ({\theta^0}_j)^{y_{ij}}(1 - {\theta^0}_j)^{(1-y_{ij})} \end{aligned} \tag{9}$$

As before, π is the success probability for the r_{kj} and S_k denotes the set of patients assigned to cluster k. Also, $p(s_1) = 1$.

From the joint density, the cluster membership depends on the prior for the cluster membership s_i, the prior for θ' and the distribution of the y_{ij}, which is Bernoulli with success probability ${\theta^\star}_{s_i,j}$.

Now the prior model for cluster membership s is proportional to either the cluster size(for existing clusters) or to a parameter α in the case of a new cluster, as discussed previously. Therefore, the posterior density for s also has two different distributions corresponding to these two cases. In the first case, for finding the posterior probability that patient i belongs to cluster k, we find that the only terms that contribute are the cluster size n_k^- and the distribution of the data. Thus, we get

$$p(s_i = k|s_{-i}, \theta^\star, r_k, y_{ij}) \propto \quad n_k^- \prod_{j=1}^{m} ({\theta^\star}_{kj})^{y_{ij}} (1 - {\theta^\star}_{kj})^{(1-y_{ij})} \quad k = 1, \ldots, K^-$$

$$p(s_i = k|s_{-i}, \theta^\star, r_k, y_{ij}) \propto n_k^- \prod_{j=1}^{m} ({\theta^\star}_{kj})^{y_{ij}} (1 - {\theta^\star}_{kj})^{(1-y_{ij})} \qquad \text{if } k = 1, \ldots, K^- \quad (10)$$

In the second case, when the patient belongs to a new cluster K, we multiply α times the distribution of the data, getting

$$p(s_i = K^- + 1|s_{-i}, \theta^\star, r_k, y_{ij}) \propto \alpha \prod \int {\theta^\star}^{y_{ij}}_{K^-+1,j} (1 - {\theta^\star}_{K^-+1,j})^{1-y_{ij}} *$$
$$[\pi \text{Beta}(\theta^\star|a_0, b_0) + (1 - \pi)\delta_{\theta^0}(\theta^\star)] d{\theta^\star}_{K^-+1,j} \text{if } k = K^- + 1 \quad (11)$$

[Note: π here is the probability that the relevance indicator $r_{kj} = 1$]

In the posterior equation above, where we consider placing the ith unit into a new singleton cluster, the success probability for the Bernoulli distribution of the data is not known. We therefore have to marginalize with respect to ${\theta^\star}_k$ for the new cluster. Recalling that $\theta^\star$ on any given attribute j can either be θ' or θ^0(it can either be a special, distinguishing rate or just the background rate), in the case that an attribute is special, we integrate out the special rate ${\theta'}_{kj}$, conditioned on the relevance vector $r_{kj} = 1$ and the special rate ${\theta'}_{kj}$ coming from the corresponding prior distribution $\text{Beta}({\theta'}_k|a_0, b_0)$. In the case that ${\theta^\star}_{kj}$ on the given attribute is just the background rate, we do not need to do any integration since the background rate is a known quantity, and we just condition on the relevance vector $r_{kj} = 0$, which contributes the $1 - \pi$ term. Thus, we need the integration only to determine the values of the special rates for the new cluster, the background rates being fixed for the current iteration of the MC sampler.

The integral in the formula for the posterior for the case when we open a new cluster $K^- + 1$ can be simplified in terms of Γ integrals. We show this below. Let $x = {\theta^\star}_{K^-+1,j}$, $x_0 = {\theta^0}_j$and $y = y_{ij}$. Then the integral on the right hand side of the equation above becomes

$$p(s_i = K^- + 1 | s_{-i}, \theta^\star, r_k, y_{ij}) \propto$$

$$\prod_{j=1}^{m} \int x^y (1-x)^{1-y} [\pi \text{Beta}(x|a_0, b_0) + (1-\pi)\delta_{x_0}(x)] dx$$

$$= \prod_{j=1}^{m} \int x^y (1-x)^{1-y} [\pi \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} x^{a_0-1}(1-x)^{b_0-1}] dx + (1-\pi)x_0^y(1-x_0)^{1-y}$$

$$= \prod_{j=1}^{m} \{\pi \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} \int x^{a_0+y-1}(1-x)^{b_0-1+1-y} dx] + (1-\pi)x_0^y(1-x_0)^{1-y}\}$$

$$= \prod_{j=1}^{m} \{\pi * \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} * \frac{\Gamma(a_0+y)\Gamma(b_0+1-y)}{\Gamma(a_0+b_0+1)} + (1-\pi)x_0^y(1-x_0)^{1-y}\} \quad (12)$$

Above, we use the fact that $\text{Beta}(x|a_0+y-1, b_0+y-1)$ integrates to 1 since it a probability density function to compute that

$$\int x^{a_0+y-1}(1-x)^{b_0-1+1-y} = \int x^{a_0+y-1}(1-x)^{b_0+1-y-1}$$

$$= \frac{\Gamma(a_0+y)\Gamma(b_0+1-y)}{\Gamma(a_0+b_0+1)} \quad (13)$$

We note here that once the new cluster is formed, we then have the problem of finding out the relevance indicators r_{kj} and the special rates θ'_{kj} for this new singleton cluster. Recall the the prior model for r_{kj} is a Bernoulli, and since the success probability for r_{kj} is a conjugate Beta(either a θ' or θ^0), the posterior is also a Bernoulli. To determine the unknown success probability in the case of the new cluster(unknown since this probability follows the special cluster rate θ'_k), we have to integrate over all possible values of θ'_k, similar to how we found the posterior of s_i by integrating over $\theta^\star_k$. The integral in this case is almost exactly the same, except that we do not include the second component in the sum of the integrals for s_i. That is, we do not need to

integrate over θ^0, since in the case of the posterior for r_{kj}, we are interested only in the success probability, θ'_{kj}. Once we have this, we may then generate the posterior values for r_{kj}. This success probability is thus exactly the first component in the sum of integrals for the posterior of s_i and is given below.

$$r_{kj} = \begin{cases} 1, & \text{with prob}=\pi * \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} * \frac{\Gamma(a_0+y_{ij})\Gamma(b_0+1-y_{ij})}{\Gamma(a_0+b_0+1)} \\ 0, & \text{with prob}=(1-\pi) * {\theta^0}_j^{(y_{ij})}(1-{\theta^0}_j)^{(1-y_{ij})} \\ & \text{for k} = K^- + 1, j = 1, \ldots m \end{cases} \tag{14}$$

Similarly, we can show that the posterior distribution of the special rate for the new singleton cluster is given by

$$\theta'_{kj} \sim \text{Beta}(a_0 + a_{kj}, b_0 + b_{kj}) \tag{15}$$

where the a_{kj} and b_{kj} will be obtained using the likelihood and updating θ'_k , explained in the following sections.

9 Conditional posterior for other parameters

9.1 Posterior for relevance indicators r_{kj}

We have already seen that it is the relevance vector that decides which of the attributes j are "relevant" or special for a given cluster k. To find the posterior density, again we use the principle that the posterior is proportional to the joint distribution(= prior x likelihood). The only terms in the joint that depend on r are the Bernoulli(π) prior for r and the probability rate $\theta^{\star}$ of the Bernoulli distribution of the data. The parameter r is involved in the distribution of the y_{ij} since it determines whether the success rate θ of the Bernoulli model for y_{ij} is special or background. Specifically, we have :

$$\theta_{ij} = r_{s_i j}\theta'_{s_i j} + (1 - r_{s_i j})\theta^0{}_j \tag{16}$$

We are interested in the posterior probability that $r_{kj} = 1$ for any given cluster k and attribute j, and this can be found by taking the product over all the patients in cluster k who exhibit attribute j(have $y_{ij} = 1$) given that $r_{kj} = 1$. This can be seen from the Bayesian conditional probability statement as

$$p(r_{kj}|y_{ij}) \propto \begin{cases} p(r_{kj} = 1) \cdot \Pi_{i \in S_k} p(y_{ij}|r_{kj} = 1), & \text{r=1} \\ p(r_{kj} = 0) \cdot \Pi_{i \in S_k} p(y_{ij}|r_{kj} = 0), & \text{r=0} \end{cases} \tag{17}$$

We have:

$$p(r_{kj} = 1|y_{ij}) = p(r_{kj} = 1) \cdot \Pi_{i \in S_k} \frac{p(y_{ij}|r_{kj} = 1)}{p(y_{ij}|r_{kj} = 1) \cdot p(r_{kj} = 1) + p(y_{ij}|r_{kj} = 0) * p(r_{kj} = 0)} \tag{18}$$

Then, for fixed cluster k and attribute j :

$$p(r_{kj} = 1|y_{ij}) \propto Ber(r_{kj}|\pi) * \Pi_{i \in S_k} \theta'^{y_{ij}}_{kj} (1 - \theta'_{kj})^{(1-y_{ij})} \tag{19}$$

Define $A_{kj} = \sum_{i \in S_k} y_{ij}$, and let n_k be the total number of patients in cluster k. Then the above equation can be re-written as:

$$p(r_{kj} = 1|y_{ij}) \propto Ber(r_{kj}|\pi) * \theta'^{A_{kj}}_{kj} (1 - \theta'_{kj})^{(n_k - A_{kj})} \tag{20}$$

Substituting for the probability of the data or the likelihood:

$$p(r_{kj} = 1|y_{ij}) = \frac{Ber(r_{kj}|\pi)\Pi_{i \in S_k} (\theta'^{y_{ij}}_{kj} \theta'_{kj})^{(1-y_{ij})}}{Ber(r_{kj}|\pi)\Pi_{i \in S_k} \theta'^{y_{ij}}_{kj} (1 - \theta'_{kj})^{(1-y_{ij})}) + (1 - Ber(r_{kj}|\pi))\Pi_{i \in S_k} \theta^{0 y_{ij}}_{j} (1 - \theta^0{}_j)^{(1-y_{ij})})} \tag{21}$$

$$p(r_{kj} = 1|y_{ij}) = \frac{\pi \Pi_{i \in S_k} \theta'^{y_{ij}}_{kj} (1 - \theta'_{kj})^{(1-y_{ij})}}{\pi \Pi_{i \in S_k} \theta'^{y_{ij}}_{kj} (1 - \theta'_{kj})^{(1-y_{ij})}) + (1 - \pi)\Pi_{i \in S_k} \theta^{0 y_{ij}}_{j} (1 - \theta^0{}_j)^{(1-y_{ij})})} \tag{22}$$

The above is the same as:

$$p(r_{kj} = 1|y_{ij}) = \frac{\pi \theta'^{A_{kj}}_{kj} (1 - \theta'_{kj})^{(n_k - A_{kj})}}{\pi \theta'^{A_{kj}}_{kj} (1 - \theta'_{kj})^{(n_k - A_{kj})}) + (1 - \pi)\theta^{0 A_{kj}}_{j} (1 - \theta^0{}_j)^{(n_k - A_{kj})})} \tag{23}$$

9.2 Posterior for θ'_k

Again we use the fact that the posterior density is proportional to the Prior x Likelihood. The prior model assigned to θ'_k was Beta(a_0, b_0) and the likelihood distribution of the data is a product of Bernoullis with success probability equal to the special rate for attribute j. Thus, the posterior density for a fixed cluster k and attribute j is proportional to the product of the prior Beta(a_0, b_0) and the product of the Bernoullis over all the members of the specific cluster k, that is, over all the patients i such that $s_i = k$ in the cluster membership function. The product of Bernoullis has success rate equal to θ'_k, the special rate or θ^0, the background rate depending on whether $r_{kj} = 1$ or $r_{kj} = 0$ respectively. Since we are interested in the posterior density for just the special rate θ'_k, we will be using only those attributes j which are relevant and follow the special rate θ'_{kj}. The attributes j that follow the background rate θ^0 contribute only to the constant of proportionality in the formula for the posterior for θ'_{kj}. We have, for a specific value of k and j and $S_k = \{i : s_i = k\}$:

$$p(\theta'_{kj}|s_i, r_k, y_{ij}) \propto \begin{cases} \text{Beta}(\theta'_k|a_0, b_0) \prod_{i \in S_k} (\theta'_{kj})^{y_{ij}} (1 - \theta'_{kj})^{(1-y_{ij})} & \text{if } r_{kj} = 1 \\ \text{Beta}(\theta'_k|a_0, b_0) * 1 & \text{if } r_{kj} = 0 \end{cases}$$

We rewrite the above using the formula for the Beta(a_0, b_0) prior as

$$p(\theta'_{kj}|s_i, r_k, y_{ij}) \propto \begin{cases} \theta'^{(a_0-1)}_{kj} (1 - \theta'_{kj})^{(b_0-1)} \prod_{i \in S_k} (\theta'_{kj})^{y_{ij}} (1 - \theta'_{kj})^{(1-y_{ij})} & \text{if } r_{kj} = 1 \\ \theta'^{(a_0-1)}_{kj} (1 - \theta'_{kj})^{(b_0-1)} * 1 & \text{if } r_{kj} = 0 \end{cases}$$

Now to simplify the right hand side above, we introduce a_{kj} and b_{kj} defined as follows. For a given cluster k and an attribute j which is relevant for that cluster, that is, such that $r_{kj} = 1$, let a_kj be the number of patients for whom $y_{ij} = 1$, and b_{kj} be the number of patients for whom $y_{ij} = 0$. Thus, a_{kj} counts the number of patients

in cluster k who actually have a symptom j which is relevant, while b_{kj} counts the number who do not exhibit the symptom j even though it is relevant for the cluster k. We can denote this as

$$a_{kj} = \begin{cases} \sum_{S_k} I(s_i = k, y_{ij} = 1) & \text{if } r_{kj} = 1 \\ 0 & \text{if } r_{kj} = 0 \end{cases}$$

and similarly

$$b_{kj} = \begin{cases} \sum_{S_k} I(s_i = k, y_{ij} = 0) & \text{if } r_{kj} = 1 \\ 0 & \text{if } r_{kj} = 0 \end{cases}$$

The product in the posterior density for ${\theta'}_{kj}$ then simplifies to

$$p({\theta'}_{kj}|s_i, r_k, y_{ij}) \propto {\theta'}_{kj}^{(a_0-1)}(1-{\theta'}_{kj})^{(b_0-1)}{\theta'}_{kj}^{(a_{kj})}(1-{\theta'}_{kj})^{(b_{kj})} \tag{24}$$

where the right hand side becomes

$$p({\theta'}_{kj}|s_i, r_k, y_{ij}) \propto \text{Beta}(a_0 + a_{kj}, b_0 + b_{kj})$$

9.3 Posterior for θ^0

The procedure for finding the posterior density of θ^0 is similar to that used for ${\theta'}_k$. Of course, here we will be interested in this patients who contribute to the background rate rather than the special rate in a given cluster k. Proceeding as before, we note that the prior model assigned to θ^0 was $\text{Beta}(a_0, b_0)$ and the likelihood distribution of the data is a product of Bernoullis with success probability equal to the special rate ${\theta'}_k j$ for attribute j in cluster k. The important difference is that θ^0 is not cluster specific, rather, it is the identical across clusters for a fixed attribute j. However, in computing the product of Bernoullis over all patients with the background

rate, we note that these are exactly those who do not follow the special rate, that is, all such patients must necessarily be in a cluster k for which the attribute j is not relevant or for which $r_{kj} = 0$. Thus, analogously with the previous case with ${\theta'}_k$, for a fixed attribute j only those clusters with the value of r_{kj} of interest, that is, $r_{kj} = 0$, contribute to the posterior density, and the remaining clusters with $r_{kj} = 1$ contribute to the constant of proportionality.

$$p({\theta^0}_j|s_i, r_k, y_{ij}) \propto \begin{cases} \text{Beta}(a_0, b_0) \prod_{i=1}^{n} (\theta^0)^{y_{ij}} (1-\theta^0)^{(1-y_{ij})} & \text{if } r_{kj} = 0 \\ \text{Beta}(a_0, b_0) * 1 & \text{if } r_{kj} = 1 \end{cases} \tag{25}$$

We can then rewrite the posterior as

$$p(\theta^0|s_i, r_k, y_{ij}) \propto \text{Beta}(a_0, b_0) \prod_{k:r_{kj}=0} \prod_{i \in S_k} (\theta^0)^{y_{ij}} (1-\theta^0)^{(1-y_{ij})}$$

whe re as before,

$$S_k = \{i : s_i = k\}$$

Again we use the counting variable a_{kj} and b_{kj} to further simplify the posterior density. For a given cluster k and an attribute j which is not relevant for that cluster, that is, such that $r_{kj} = 0$, let a_kj be the number of patients for whom $y_{ij} = 1$, and b_{kj} be the number of patients for whom $y_{ij} = 0$. Thus, a_{kj} counts the number of patients in cluster k who have a symptom j which is not relevant, while b_{kj} counts the number who do not exhibit the symptom j. We can denote this as

$$a_{kj} = \begin{cases} \sum_{S_k} I(s_i = k, y_{ij} = 1) & \text{if } r_{kj} = 0 \\ 0 & \text{if } r_{kj} = 1 \end{cases}$$

and similarly

$$b_{kj} = \begin{cases} \sum_{S_k} I(s_i = k, y_{ij} = 0) & \text{if } r_{kj} = 0 \\ 0 & \text{if } r_{kj} = 1 \end{cases}$$

The product in the posterior density for θ^0 then simplifies to

$$p(\theta^0 | s_i, r_k, y_{ij}) \propto \theta^{0(a_0 - 1)} (1 - \theta^0)^{(b_0 - 1)} \theta^{0(a_{kj})} (1 - \theta^0)^{(b_{kj})} \tag{26}$$

where the right hand side becomes

$$p(\theta^0 | s_i, r_k, y_{ij}) \propto \text{Beta}(a_0 + a_{kj}, b_0 + b_{kj})$$

10 Results

I initially simulated the data,as already mentioned, along with their cluster membership under the prior model(the simulation truth). To look at how well the MCMC performed, I plotted the posterior distribution of the number of clusters, over 2000 iterations. The mode of this distribution is similar to the true number of clusters in the simulated data. The results for the average value of the special and background attribute rates are not so good; the values are not always close to the simulation truth.

Figure 1 shows posterior distribution $p(K \mid y)$. Recall K is the number of clusters.

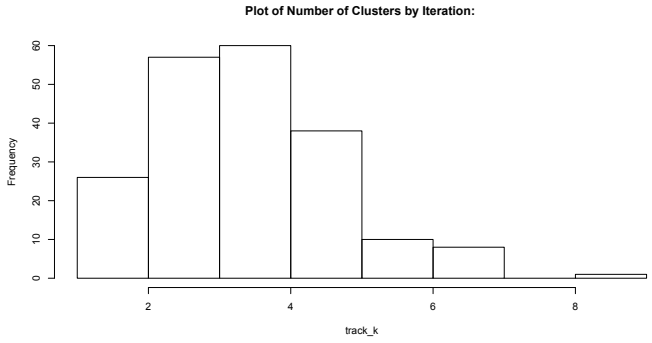

Figure 1: Disribution of number of clusters over 2000 iterations

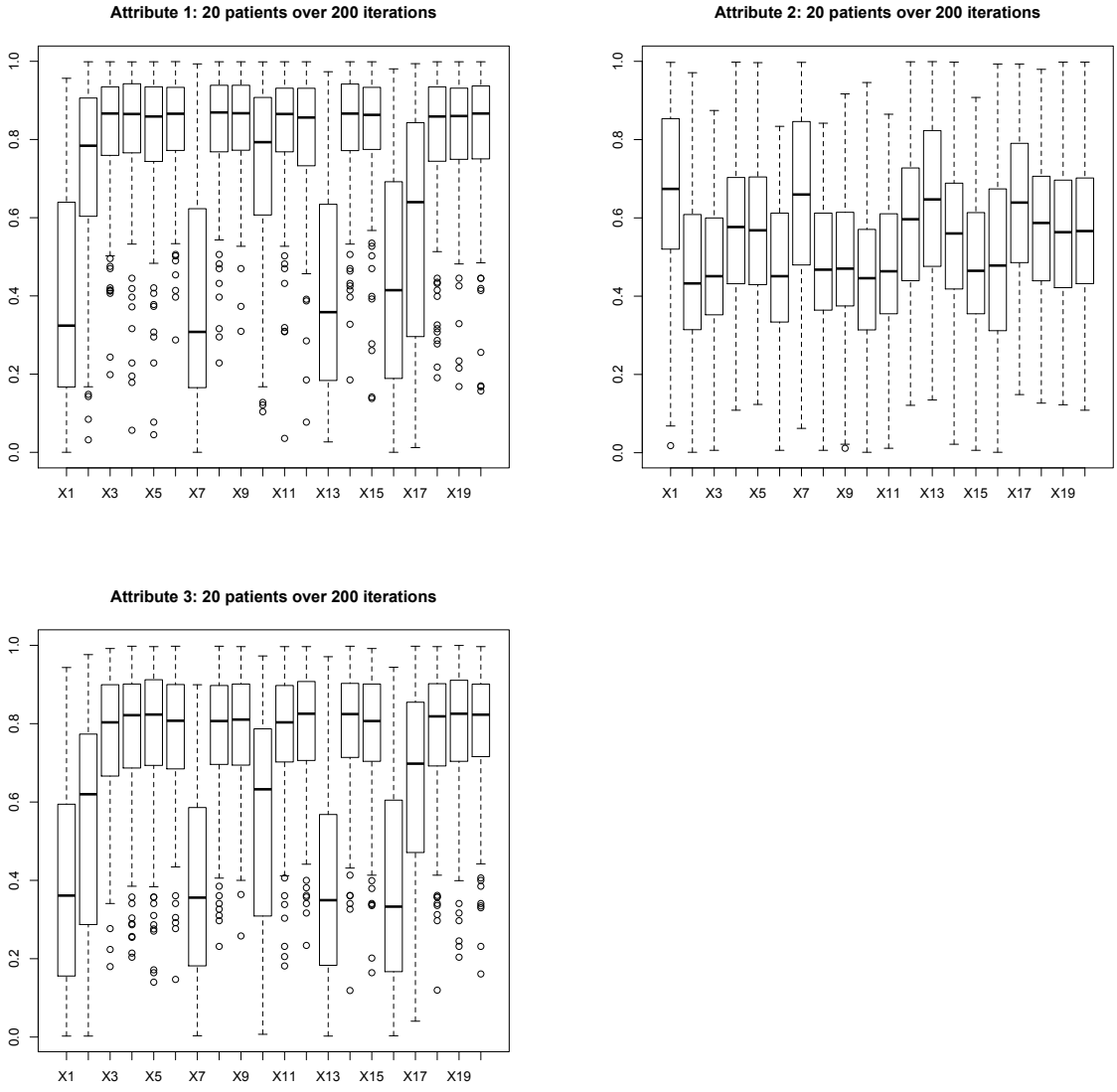

Figure 2: Posterior Distributions of Attributes 1,2 and 3

11 Conclusion

I discussed some techniques for clustering binary data in this report. These techniques can also be examined on binary data from other sources, such as in data on loans and credit ratings, where we have information on the customers' profile which is based on the presence or absence of some attributes. The difference that we might have to consider in these cases is the size of the data might be larger, and we might have to design the MCMC to account for the size of the data.

References

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: Wiley.

A. K. Jain and R. C. Dubes(1988), "Algorithms for Clustering Data", Prentice Hall

Hoff, P. D. (2005). Subset clustering of binary sequences, with an application to genomic abnormality data, *Biometrics*, 61, 1027-1036

Hoff, P.D. (2009). *A First Course in Bayesian Statistical Methods*, Springer Verlag, New York.

Steinley, Douglas (2006). K-means clustering: A half-century synthesis *British Journal of Mathematical and Statistical Psychology*, 59, 1-34